DOCUMENT RESUME

ED 276 741 TM 860 690

| | |
|---|---|
| AUTHOR | Anderson, Patricia S. |
| TITLE | Beyond the Wall Chart: Issues for States. |
| INSTITUTION | Northwest Regional Educational Lab., Portland, OR. Assessment and Evaluation Program. |
| SPONS AGENCY | Office of Educational Research and Improvement (ED), Washington, DC. |
| PUB DATE | 30 Sep 86 |
| CONTRACT | 400-86-006 |
| NOTE | 50p. |
| PUB TYPE | Reports - Research/Technical (143) |
| EDRS PRICE | MF01/PC02 Plus Postage. |
| DESCRIPTORS | *Academic Achievement; Accountability; Achievement Tests; *Comparative Testing; Costs; Data Collection; Educational Assessment; Elementary Secondary Education; Equated Scores; *Evaluation Problems; National Programs; Program Implementation; Research Needs; *State Programs; *Testing Problems; *Testing Programs |
| IDENTIFIERS | Alaska; Council of Chief State School Officers; Department of Education; Hawaii; Idaho; *Indicators; Oregon; Washington |

ABSTRACT

The Council of Chief State School Officers has raised some issues pertinent to the implementation of a national project to collect statistical indicators of academic achievement and to compare them across states. The United States Department of Education shares this concern, with an emphasis on sampling schools for a school level analysis which may be aggregated to the state level. Three issues raised by the Chiefs include the identification of subject matter domains for assessment, the scale for analysis and reporting of results, and the administration and standardization of testing across the states. Potential problems with national testing include the following: (1) redirection of state or local curriculum goals; (2) high costs of a better method of cross-state comparisons than the current Wall Chart; (3) separating the effects of student differences from curriculum or teaching effects; (4) difficulty in making useful conclusions from long-term testing; (5) conflict between state and local data indicators; (6) selection of a model which does not restrict data collection and analysis; and (7) practical implementation problems. Testing programs in five Northwestern states are briefly summarized: Alaska, Hawaii, Idaho, Oregon, and Washington. Montana has no state testing program. (GDC)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED276741

# BEYOND THE WALL CHART: ISSUES FOR STATES

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Kirkpatrick

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Patricia S. Anderson
Principal Author
Northwest Regional Educational Laboratory
Evaluation and Assessment
300 Southwest Sixth Avenue
Portland, Oregon 97204

September 30, 1986

This paper is based upon work performed pursuant to Contract 400-86-006 of the Office of Educational Research and Improvement. It does not, however, reflect the views of that agency or Northwest Regional Educational Laboratory.

BEST COPY AVAILABLE

TM 860 690

BEYOND THE WALLCHART: ISSUES FOR THE STATES


| | | |
|---|---|---|
| Introduction | | 1 |
| Part I. | Assessment Programs In the Northwest States | 4 |
| Part II. | The Promise of National Indicators | 6 |
| Part III. | Issues in Implementing Performance Comparisons | 12 |
| | 1. State Or Local Goals May Be Redirected | 12 |
| | Restructured State Assessment | 13 |
| | Redirected Curriculum | 15 |
| | Educational Outcomes | 16 |
| | Misplaced Direction | 16 |
| | Setting Goals | 17 |
| | 2. Costs Could Displace Existing Educational Programs | 18 |
| | 3. Differences Among States Could Distort Educational Differences | 21 |
| | 4. Long Term Utility Of National Comparisons Present Problems | 22 |
| | 5. State vs Local Indicators: Practical and Statistical Problems | 23 |
| | 6. The Model and The Measures Potentially Restrict The Collection and Analysis of Data | 25 |
| | The Model of Education | 26 |
| | The Measures | 27 |
| | 7. Practical and Logistical Problems Threaten Accuracy | 31 |
| | Test Cycle | 31 |
| | Reporting | 32 |
| | Sampling Frame | 32 |
| | Exclusions | 34 |
| | Standardization of procedures | 35 |
| | Data Burden | 35 |
| Executive Summary | | 37 |
| References | | |
| Appendix | | |

BEYOND THE WALL CHART: ISSUES FOR THE STATES

**Background**

The flurry of reports on the decline of education in America has catapulted the use of statistics to the top of the educational agenda as a means of ensuring that the many recommendations for reform will, in fact, improve educational attainment. This interest in statistics or indicators of educational achievement is now a public focus because of the demand, particularly at the federal level, for educational accountability and the acknowledged need for timely, accurate, relevant and valid measures of achievement. This paper will lay the foundation with a description of the national efforts to develop cross-state comparative indicators. It will also review activities in the six Northwest state assessment programs, outline the promises for the national program and suggest issues for states.

The National Commission on Excellence in Education Task Force who wrote A Nation at Risk recognized the great difficulty in obtaining indicators of educational achievement. Other reporters on the status of education also lamented the "sorry state of educational statistics." Interest in national statistics on education is not limited to the Office of Education where almost two years of study of "The Redesign" of educational statistics is under consideration. The Council of Chief State School Officers (CCSSO) has called for a collaborative effort to design a cross-state system of indicators for educational accountability. In 1984, the CCSSO voted to move forward with their project that would culminate in comparative educational indicators. For over a year, the National Governors' Association has been working on recommendations for a five-year education agenda. The chairman of that task

force, Governor Lamar Alexander of Tennessee, has recently been selected by Secretary of Education William J. Bennett to serve as chair of a joint U.S. Department of Education and National Academy of Education Task Force on ways to improve measurement of the performance of students and schools across the country.

It is clear from this activity that the governors, legislators and Chief State School Officers--who control the majority of educational expenditures—are increasingly being expected to provide proof of educational attainment to the general public and active business coalitions. The wave of enthusiasm for numbers—ratings and rankings--has met with abbreviated accounts and at times, equivocation. Thus, there is some feeling by elected officials and the public that professional educators have deliberately withheld information or have provided data that is meant to obfuscate. When cars and refrigerators can be rated and compared, the public does not readily understand or accept why it is not possible to measure and report educational outcomes.

The major task in achieving national educational accountability is to decide what common measures to collect and how to collect them. Decisions on these issues will impact who will pay for it. The proposals at the national level have so far centered on the development of common indicators of educational achievement and the concomitant indicators that will describe or explain any differences that might exist. Attention has been given to methods of data collection (e.g., sampling issues, existing versus new data collection). The impact and interaction of national efforts with state and local efforts has not yet been thoroughly considered. "Who will pay?" is another question that appears not to have been clearly and explicitly addressed.

The purpose of this paper is to provide information on the issues surrounding the development of cross-state comparative indicators to the Chief State School Officers and others in the six member states of the Northwest Regional Educational Laboratory. These issues can be broadly defined as policy issues and technical issues. Policy issues primarily focus on the desirability of the proposal while the technical issues focus on feasibility. These, however, are not mutually exclusive. In many instances issues may overlap with policy issues often giving guidance to technical ones.

In this discussion, it is not the purpose to endorse or contrast either of the two plans for a national indicator project: the Chiefs' Evaluation and Assessment Center project, or the U.S. Department of Education's plan for the redesign of elementary and secondary data collection. Rather, they will be considered as one. Both have a similar goal: to collect comprehensive indices of educational performance across the states. They differ, however, in their focus and procedures for developing their plans. The Chiefs' project is currently designed to collect data samples for inter-state analysis. The U.S. Department of Education is sampling schools for a school level analysis, which may be aggregated to the state level.

This paper will complement the Chiefs' recent White Paper (April 30, 1986) on options for three issues pertinent to implementation of a national indicator project: (1) identification of subject matter domains for assessment; (2) the scale for analysis and reporting of results and (3) administration and standardization of testing across the states. Comment will be made on these as well as the broader policy issues such as cost, reorganization of state and local testing goals, and narrower, but perhaps more intractable issues, such as developing common definitions.

Part I will briefly identify the existing assessment programs in the six Northwest states. Part II will present the goals or promises of a national indicator program. Part III will identify the issues related to implementing such a program. Thus, we will examine where we are, where we are going, and the issues surrounding how we are going to get there.

Finally, each issue raises the prospect of a decision that will be made. States may participate in making these choices as they consider their decision to participate in the national project. States now have an opportunity to impact the design by raising issues that are of concern. A brief listing of these decisions are included in the appendix.

## PART I. ASSESSMENT PROGRAMS IN THE SIX NORTHWEST STATES

Five of the six Northwest states have statewide assessment programs. Salient features of these programs are presented in Table 1. Most noteworthy are their differences. Variation occurs in the testing instruments, age levels tested, testing cycle, number and choice of grade levels tested and cost per student.

---

Refer to Table 1

---

Similarity occurs in the purposes or goals of the testing program. All of the states with assessment programs use the results for curriculum improvement and public accountability. In this regard, assessment directors revealed that

TABLE 1

NORTHWEST STATE ASSESSMENT CHARACTERISTICS

| State | Test | Subjects | Grade | Cycle | Population | Purposes | Cost (Per student) |
|---|---|---|---|---|---|---|---|
| Alaska | State Developed | Math<br>Reading | 4, 8<br>4, 8 | Bi-annual<br>Winter | Universe<br>@ 15,000 | Curriculum improvement<br>Public accountability<br>Student diagnostic * | $50-60,000<br>($4/student) |
| Hawaii | SAT | Math<br>Reading<br>Writing | 3, 6, 8, 10<br>3, 6, 8, 10<br>3, 6 | Annual<br>6, 8, 10 Fall<br>3 Spring | Universe<br>@ 50,000 | Curriculum improvement<br>Public accountability<br>Student diagnostic * | $200,000<br>($4/student) |
| | CBM | Basic Skills | 3 | | | School level improvement | |
| Idaho | ITBS., TAP, State Developed | Math<br>Reading<br>Writing<br>Science<br>Social Stud.<br>Wrtg/ref skls | 8, 11<br>8, 11<br>8, 11<br>8, 11<br>8, 11<br>8, 11 | Annual<br>Winter | Universe<br>@ 30,000 | Curriculum improvement<br>Public accountability<br>Student diagnostic * | $46,000<br>($1.53/student) |
| Montana | No state Testing Program | | | | | | |
| Oregon | State Developed | Reading<br>Math<br>Writing | 8<br>8<br>8 | Annual<br>Winter | Sample<br>@ 25,000 | Curriculum improvement<br>Public accountability | $100,00<br>($4/student) |
| Washington | MAT | Math<br>Reading/lang | 4, 8, 10<br>4, 8, 10 | Annual<br>Fall | Universe<br>@ 110,000 | Curriculum improvement<br>Public accountability<br>Student Diagnostic * | $150,000<br>($1.36/student) |

Table drawn from interviews with state Assessment and Evaluation staff, February and June 1986.

*Student diagnostic should be interpreted to include placement or selection for special programs as well as diagnostic information to plan individual students' specific instruction in a classroom.

data collected over time is viewed as one of the common strengths. Their trend data are used for reviewing curriculum changes and the impact of various reforms.

Four of the states use the state testing program for individual student testing which may provide supplemental school placement decisions and parent information. Providing individual student test results was viewed as a service to smaller districts and necessitated a move from sampling to universe testing.

Variation within the states is as great or greater than that among the states (Coe, 1986). Local educational agencies report individual diagnostic testing for placement and remediation, curriculum development and assessment, and reporting to parents. Many times, the local school districts use tests other than those used by the state, and they test at other grade levels and times. Some local agencies, however, report little or no additional assessment activity, while others report extensive local testing. Thus, with the exception of math and reading tests in all the five states, there is little commonality in the Northwest state assessment programs in number of subject areas, time of year for testing, grade levels tested or specific tests used. This suggests some difficulty in using existing state assessment programs for any interstate comparisons.

## PART II. THE PROMISE OF NATIONAL INDICATORS

The goals of a national assessment of educational performance include accounting for the use of fiscal resources; comparing the effectiveness of state or district reforms; ensuring economic advantage; generating public support; balancing scarce resources among the states through federal assistance and standardizing the curriculum. These are discussed below.

The promise for using educational indicators is that they will help educators at all levels be accountable for the vast amount of money spent on education, and they will assist states and districts in sorting out productive reforms. To that end, national indicators of achievement have already been produced in the form of the "Wall Chart" first prepared by the U.S. Department of Education in 1985. The Wall Chart presents states educational statistics such as average test scores for the Scholastic Aptitude Test and the American College Test, and a number of contextual factors such as degree of poverty. The purpose of these comparisons was to present to the public "for the first time" a concept of academic accountability that went beyond counts of students, teachers, courses or costs. The Wall Chart offers one comparative picture of the educational outcomes in each state. Its publication caused great concern among educators and researchers alike over its essential validity, reliability and fairness. It is these issues that activities surrounding the development of alternatives addresses.

Providing comparative data across states or districts is intuitively a reasonable notion. It has been driven in part by the growing movement away from the local funding of education to state funding and influence. Figure 1 graphically presents these shifts . Figure 2 indicates these shifts as well as the proportion of federal, state or local funds for education. In each of the Northwest states, the percentage of state funding has increased from a decade ago. One consequence of assuming increased financial responsibility, is for states to set directions and standards for local educati

The promise for using educational indicators is that they will help educators at all levels be accountable for the vast amount of money spent on education, and they will assist states and districts in sorting out productive reforms. To that end, national indicators of achievement have already been produced in the form of the "Wall Chart" first prepared by the U.S. Department of Education in 1985. The Wall Chart presents states educational statistics such as average test scores for the Scholastic Aptitude Test and the American College Test, and a number of contextual factors such as degree of poverty. The purpose of these comparisons was to present to the public "for the first time" a concept of academic accountability that went beyond counts of students, teachers, courses or costs. The Wall Chart offers one comparative picture of the educational outcomes in each state. Its publication caused great concern among educators and researchers alike over its essential validity, reliability and fairness. It is these issues that activities surrounding the development of alternatives addresses.

Providing comparative data across states or districts is intuitively a reasonable notion. It has been driven in part by the growing movement away from the local funding of education to state funding and influence. Figure 1 graphically presents these shifts . Figure 2 indicates these shifts as well as the proportion of federal, state or local funds for education. In each of the Northwest states, the percentage of state funding has increased from a decade ago. One consequence of assuming increased financial responsibility, is for states to set directions and standards for local educational agencies. Information may be collected on how these directions are followed and what effect variations in educational choices created by state standards, laws and curricula, have on such things as teacher selection and quality and, most importantly, students.

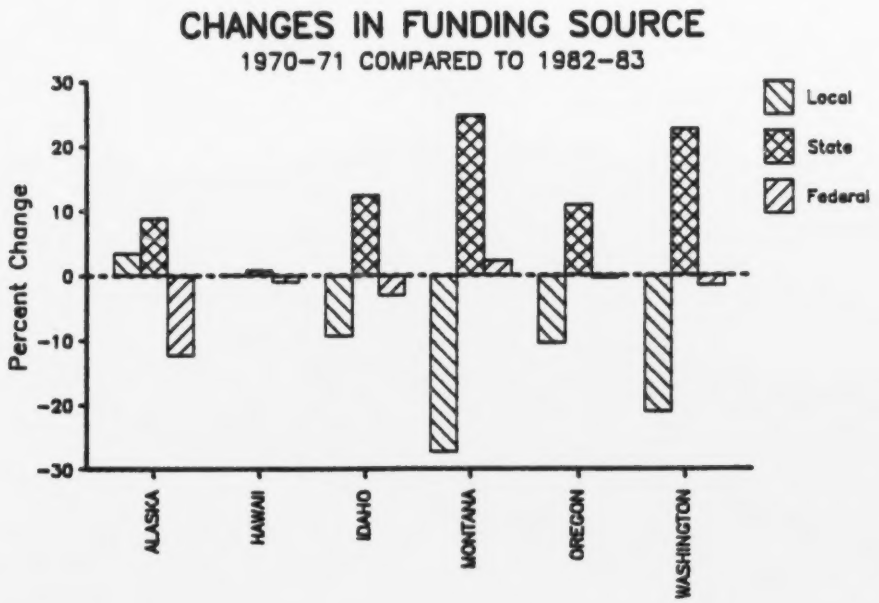


Figure 1. Increase or decrease in funding by source in the six Northwest states

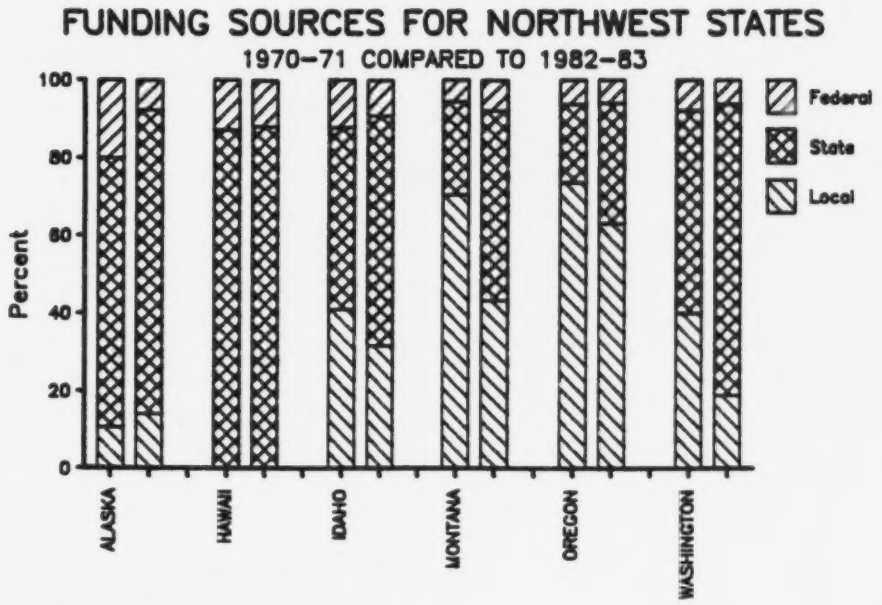


Figure 2. Percent of funds by source for two time periods.

Making comparisons is not a new phenomenon. Parents and communities have never been satisfied to know that their child or school had a score of "35" on a test. They want to know whether that score was "commendable" or "lower than acceptable." With the emphasis on economic competition among communities, states, and nations, the general public seeks to be assured that children in this class, school, district and state are doing as well or better than others. Not as explicit, but also a critical need, is for the public to know whether students are doing "well enough" to compete in a variety of arenas, e.g., technical, academic, military.

One potential benefit of state comparisons of educational performance, that may ensure participation and cooperation among many states, is that the favorable academic comparisons support the positive and successful bids for industrial locations. The economic motive has a tendency to separate educators, who are motivated by a desire to know what works, and public officials, who are motivated by the "bottom line." The transition from an industrial economy to a service and information economy has created great job displacement and migration from state to state. With thousands of jobs being lost to foreign nations, states are competing with each other for new business and industry. Most see education as inextricably tied to economic development and view a rush to rate states educationally as a necessary marketing tool to woo businesses. Many businesses are making location decisions based on the skills and capabilities of the local workforce. States that are willing to share their educational successes will be viewed more favorably, particularly if that accountability is believed to drive continuous analysis and revision that can result in improved education.

A related purpose for providing state and local comparisons of educational performance is to elicit the necessary public support for education in an era

when 60 percent or more of the voters do not have children in school (Cardenas and First, 1985; Moynihan, 1986). This non-school related population is less and less accepting of altruistic motives as a basis of support for education. The baby boom population of middle-aged adults is increasingly aware that if they do not support education, their own retirement security will be threatened. They are becoming painfully aware that a social security program that began with ten workers for every beneficiary, will only too soon have three workers to every beneficiary (The Futures Group, 1984). They are becoming aware that the American economy can no longer withstand a workforce in which one-half of new workers lack skills in spelling, grammar, and computing, nor can the economy support the 20 to 30 percent of the adult population not "functionally competent" to manage their own lives (Snyder, 1984). It is increasingly apparent to this generation that failure to educate the younger generation will not only affect that generation, but the following one as well.

Whatever the flaws and problems inherent in rewarding either good or poor performance (i.e., providing financial aid to those with low scores), some believe that federal financial support (or sanctions) to support school reforms will result from sharing national data with state comparisons. Just as the federal government took on much of the costs of educating handicapped youngsters (Madaus, 1979), a new federal role might be focused on school improvement and reform support.

Another purpose or promise of such comparisons is that they will provide to educators and the general public an understanding of the success or failure of the various educational reforms. Each state and district has launched its own review and is selecting from the array of reforms, those which appear to fit its people and its pocketbook. If reforms can be linked to measurable

gains or advantage in its students, others will want to emulate those reforms. The overall result is expected to be more effective education.

A more practical promise for comparisons among states or districts is that when the reforms and research produce results, curricula and teaching quality both within and across states will standardize, or at least become more public. This will allow children to move from state to state with minor disruptions in their education.

Beyond comparisons at the state level lie the potential and promise of districts (or schools) as the unit of analysis. It could be argued that the closer the unit of aggregation to students the more accurate will be the results. This is true not only in the statistical sense but also in the practical sense. Administrators who know that their school/district is being measured will be interested in the results and the quality of data reported. The publicity surrounding such reports is believed to provide a strong incentive for school or district leaders to focus on areas of deficiency while maintaining areas of strength.

Factors involved in educational attainment are often better understood by the general public when describing the local school or local district. The local people are more familiar with local factors that either raise or lower "average scores." The local communities know that they can influence and control their district/school personnel if school performance is not deemed acceptable. The human and resource factors associated with state averages are sometimes less well understood. Thus, a further promise of comparative educational performance is that local communities might be brought into the problem solving and decision making about educational performance.

District and/or school comparisons offer much promise because of the research used and cited in the effective schools movement. This movement is based on the premise that it is the school where learning and decision making about learning take place. Accordingly, when comparisons of schools provide the basis for how a school is doing related to others, both high- and low-achieving schools provide a reference for further analysis. Successful efforts can be emulated and unsuccessful ones discarded.

Taken together, the promises of state (or local) comparisons of educational success present as an opportunity to establish, maintain and promote support for education as a legitimate public expenditure in the forefront of the public agenda.

## PART III. ISSUES IN IMPLEMENTING PERFORMANCE COMPARISONS

The 1984 vote of the Chiefs to move toward state comparisons established a clear policy direction. This vote was not taken lightly and reflects the political imperative of the majority of the chiefs. Yet in both designing and implementing such a process, significant issues must be addressed. Issues include redirecting state or local goals, structures, cost, usefulness, state vs. local indicators, measures and measurement framework, and specific implementation issues such as data burden, testing cycle, sampling and exclusions.

### 1. State Or Local Goals For Education May Be Redirected.

Redirecting the goals of education or its organization presents significant issues to states and districts. Issues discussed are the organizational and structural issues of a state office of assessment, changing

curriculum or outcomes of education and setting goals based on test scores or other indicators. These are discussed below:

Restructured State Assessment. The goals and functions of education are frequently embodied in the structure and organization of the state departments of education. One major function in five of the Northwest states is the state assessment program. Yet it appears that one of the major assumptions underlying much of the discussion and developments to date is that national indicators can be developed and collected without sacrificing existing state assessment programs. States will, however, find that it may be necessary to address the ways that the national indicators might be collected as part of, or in conjunction with, ongoing state assessments and data collection.

The "White Paper" developed by Ramsay Seldon on April 30, 1986, outlined three options for establishing the content of the assessment tests and reported that most states leaned toward a process that would use an "optimal consensus" approach. In this scheme, the content of the tests consists of what is generally agreed to be the domain that "should" be taught. It was noted that this will give states flexibility in determining what their specific emphases will be. This is considered a disadvantage by some because it will be difficult to reach agreement on subject-matter that will stretch beyond current practice or a minimum competency. In addition to reaching agreement on this issue, it is not clear how the balance will be struck between "having flexibility" to decide on emphasis and states discounting results because "we emphasize something different."

If the test content "represents the maximum breadth and depth in each subject," then it is likely that this will lead to a test with multiple forms that measure a number of dimensions, rather than a single form that is

typically used in Northwest state assessments. This is similar to the current National Assessment of Educational Progress (NAEP) process and could result in duplicated efforts, if it is conducted separately from existing assessments.

The issue of whether to have a single score that covers a broad area, e.g., reading, or whether to have multiple scores, e.g., reading comprehension, word attack, depends on who needs the information. If the assessment is only to provide state comparisons on broad subject areas rather than for making decisions about what to do, then the single measure becomes a more viable method. Ultimately, however, legislators, state boards and the public will want to know what can be done if their state is one of the 50 percent that is below the midpoint. The top 50 percent will be faced with the question of how to improve.

One of the primary reasons for a national achievement indicator is to have a comparable index that can be used to report outcomes. It is very likely that the status reported on the national indicator(s) will differ from the status reported by the state assessments. The reasons for such discrepancies can range from different test contents to different reference groups. Many states use nationally norm-referenced tests and report status in terms of average scores, percentiles or grade equivalents. Experience in California and in other states suggest that confusion and consternation can result when the public is told that we are at the XXth percentile on our local test but are at the YYth percentile on state norms. If the national indicator process results in such confusion, then one of its primary purposes, i.e. to have a common index across states that the public will readily understand, will be compromised.

States will also have to deal with the possibility that any state comparison approach will lead to yet more information in frequently assessed areas of reading and math with only secondary priority for areas such as science and social studies. If this is true, then the goal for the indicators to stretch current practices and minimum achievement levels will also be in areas for which much information already exists and will be minimal in other areas of school emphasis.

Redirected Curriculum. When traditional measures of achievement in math, reading, science, writing or the arts, are used as indicators of outcome, the curriculum may narrow to these topics and thus be restructured for all students. Focusing on academic units has the potential to restrict the variety of school experiences and even limit the array of skills appropriate for training. Thus, some schools, districts and states that have placed a heavier emphasis on vocational education, rather than academic preparation, will find their educational goals redirected.

Increasing academic requirements may be based on a mistaken assumption that all students can benefit from a more rigorous curriculum. While much has been learned about effective schooling for the disadvantaged, it is not yet clear whether the increasing numbers of disadvantaged students will be better served or achieve more through additional coursework requirements or through alternatives that match their skills and aptitudes.

Curriculum sequence may be affected when tests on certain topics are included in various grade levels. For example, if an 11th grade math test includes significant measurement of algebraic concepts, then local curriculum designers may encourage offering the basic course at the 10th grade to increase average scores.

One possible positive effect of the interest in educational accountability is the potential for more common curricula across states and districts. While a standardized curriculum offers some beneficial or promising effects, it may also result in some damaging effects if a state or district revises its goals and methods to focus on the common curriculum. If there are true differences between the students and education offered in each state, then standardizing the curriculum would have to be done with the lowest achieving students in mind, lest these students be unable to profit from the "new curriculum."

Educational Outcome. Closely related to coursework redirection, is the potential for narrowing the educational focus to only the outcomes tested or measured by any indicator program. Such redirection occurs over time, for example, as the measurement of basic skills supersedes higher order skills, or perhaps when testing reading supersedes writing.

Misplaced Direction. If the public maintains its interest in educational indicators, education at all levels may run the risk of emphasizing and acting on factors that are only mildly associated with achievement. For example, class size is viewed as an important "input/resource," "policy/practice," or "process" variable that influences achievement. If achievement is low, a state or district may conclude that reducing class size will have the desirable result. Research, however, is equivocal on the linear relationship between class size and achievement gains. Such an emphasis would be extremely costly and would likely lead to little gain. On the other hand, using a finding that no relationship exists or that class size at least is not associated with low achievement could lead to an equally indefensible policy to increase class size.

Setting Goals. Another little-explored issue is that current achievement tests were not designed to be used to set test-oriented goals. They are designed to sample those skills or knowledge generally thought to be important. The purpose of setting goals is to improve learning through appropriate expectations, incentives and, perhaps, even rewards. Using current tests invites and drives schools to focus on only those things tested--and not on the objectives, skills or even test items from which the test was developed.

One of the more difficult technical issues related to developing district or state indicators is that of using the indicators to set performance goals. Encouraging annual increases in absolute as well as relative performance, supposes that the whole range or significant portions of students can have learning gains and perform better each and every year. These gains are considered against last year's class of students. This year's fifth graders must learn and achieve more than last year's fifth graders. Next year's must do better than this year's. However, achievement score gains can be confused with learning when efforts are made to: (1) develop more "aligned" instruction, (2) provide motivational incentives or (3) assist students in becoming more testwise.

A frequent problem mentioned in current educational reform literature is the potential to increase the number of dropouts by requiring higher levels of performance or greater numbers of academic subjects to advance from grade to grade. Research on the potential impact on school retention has been equivocal. Some have argued that educators can now use the effective schooling techniques with lower performing students and achieve more favorable outcomes. They argue that lower performing students can meet higher expectations and thus will improve overall school averages. Yet, recently, the

majority of principals surveyed in Texas attributed a 14 percent increase in student dropouts to the tightened curricular and graduation requirements (Education Week, May 21, 1986).

Beyond the issues mentioned above are the practical difficulties for a local school district. Some states now use regression analysis to indicate to districts how much their scores must improve for the state average score to reach its goal. Indicating a district's "share" of state improvement could infuence a state or district's willingness to volunteer to be included in a national indicator process at best or result in local manipulation in the proctoring, scoring or selection of students, at worst.

2. Costs Could Displace Existing Education Programs

The provision of state-by-state comparisons has been accomplished through the publication of the Wall Chart. Those who desire a more valid, fair and reliable measure for comparison recognize that their enthusiasm may diminish when the costs are totaled. At this time, consideration has focused on few options:

1. Use adjustments for state demographics and percentages of students taking the SAT.
2. Develop links (e.g., anchor tests; bridge questions) between existing state tests.
3. Expand NAEP testing samples for state comparisons.
4. Support the Chiefs' project to develop commonly agreed upon indicators and procedures for collecting and reporting at the state level.
5. Support the U.S. Department of Education effort for district level sampling and reporting.

Each of the options represents challenges to utility, fairness and validity---probably in decreasing order. Each also represents additional costs to be borne by states---in ascending order.

The current Wall Chart represents the least expenditure and effort by any level of government. Adjustments or corrections for the selection of students and demographic factors must be carried out by those familiar with statistical procedures and aware of the threats to validity. Given the concerns over adjustments (Weiner, 1986), the practical matter, however, is that after scores are released, newspaper reporters see no further newsworthiness in the corrections which may come at a later date.

For the Northwest states, all of whom present respectable average SAT scores, the motivation to participate in alternative strategies is not driven by the need to equivocate on local scores to the public, but by a need to report more fairly, e.g., to control for the degree to which only a higher level or subset of students take the tests.

Developing nationally standardized items for inclusion in state assessments poses an alternative cost effective approach. This would require the appropriate equating studies such that all levels of potential difficulty could be included in the common scale. This alternative would also require states to include the appropriate questions at the various levels of difficulty in the state testing program. States with their own tests may not find these additions overly burdensome, while those with standardized publisher-developed tests would have to adjust preprinted, standardized tests to include the items. Such an item pool concept using NAEP as a basis is under consideration by CCSSO and the U.S. Department of Education.

Expanding NAEP for statewide comparisons, such as in the (Southern Regional Educational Board (SREB) pilot study, provides the state averages at a cost of $35,000 per grade per subject per state.

The Chiefs' project planning envisions a limited NAEP type assessment per subject and grade level. A minimum of $35,000 has been suggested for the one grade, one subject version. This is the cost for contractor services. States must assume responsibility for collecting the data, identifying student sample frames, coordinating the collection effort and so on. The maximum per state with three grades and five subjects is projected to be $500,000. If new data collection, such as collecting common dropout statistics, measures of teacher performance or student "engagement" were advanced, additional costs would be projected for each participating state.

The Department of Education program has not determined the cost as yet because the decision on the test design and consequent sampling design will affect the totals. Current estimates are $35 to $50 per student per subject.

The cost issue is highly related to purpose. If states/districts believe it is important to have a more fair, valid and reliable measure of state school achievement, then the cost per year per grade could easily be $160,000 plus, per state. A comprehensive NAEP type testing envisioned by at least one of the national designs, using several grade levels, would increase the costs up to $600,000 per state. It is important to note that estimated costs for these options are in addition to costs for ongoing state assessment programs. In the NWREL region, only Hawaii has state costs exceeding the most limited estimates. If the states are not able to find additional support, either through their own appropriations, or federal designations, then it might be necessary to consider either linking or equating state assessments or making the national assessment a subset of state assessments.

With increasing demands from all public spending areas in the states, and budgetary ceilings in many, monies for any testing program are likely to come from existing state and federal budgets. State legislators may find it difficult to revise spending priorities for educators who need to know "what to do about it" when all the public wants to know is "who won."

3. Differences Among Students Confound Educational Outcome Differences

The use of comparative educational indicators implies that the factors associated with differences among states in educational performance may be attributable to differences in state curriculum or teaching quality. These factors, however, are confounded by numerous other "contextual" factors that differentiate the states as well. These include demographic factors such as the number and percentage of minority students, low income students, students whose primary language is not English, exceptional students (both gifted and special education) and resources within the state available for education (e.g., assessed valuation, per capita wealth, per student spending). Since it is well known that these demographic variables are associated with achievement scores (White, 1982;Powell and Steelman, 1984), differences in test scores across states are due to the variable distribution of populations and resources across the states rather than to the quality of education alone. These differences must be statistically adjusted across states in order to assess the effects of the educational system alone. There are, however, technical problems associated with this adjustment. But more importantly, the public wants to know why all students cannot equally profit from the educational opportunities offered by the states. Adjusting has the effect of removing from consideration the scores for any selected group (e.g., lower socioeconomic, limited English) in order to test the effects of education alone.

If one of the reasons for having indicators of educational performance is to suggest to the policy makers and potential businesses that a given state's population has an acceptable achievement level compared to other states then adjusting for differences due to populations only obscures the differences that the business will find if it moves to that state seeking to hire that "average" person.

A common solution to adjusting is to present data from various subgroups separately. Thus, the NAEP or SREB will present average scores for white and black students. Such a presentation may urge educators to focus on the needs of specific groups, but it may also serve to magnify prejudice or reduce efforts because of perceived futility.

**4. Long-Term Utility of National Comparisons Presents Challenges**

Another issue associated with comparing state averages is the ultimate utility of the measures offered. Comparisons offered on norm-referenced tests suggest that a number of states will be above the average and a number will be below. Under a hypothetical condition of equal score distributions in each state, average test scores (and ranking) could be "low" one year and "high" the next year because of chance (sampling error) alone. It should be expected that the average scores among states could fall on a bell shaped curve. If that happened, the majority of states or districts would fall in the middle, where the distinctions become so fine that they lose all meaning.

Criterion-referenced tests could offer equally misleading results. If state variation in test performance was large, then the test would have to be developed with the lowest performing states in mind, lest too many of their students fail. If an "easy" test were created, 100 percent of students in other states might pass.

One of the arguments for national comparisons is that educators and the public alike will be able to determine "what works" by examining the educational systems of high- and low-achieving states. Given that most, if not all, of the states are introducing many reforms, the analysis of the effectiveness of any given reform could easily be lost without further evaluation studies. The indicators, as currently described, would provide minimal insight.

Long-term utility of state indicators of educational performance may cause great initial concern among those states doing poorly. If reforms and greater resources do not produce results, then apathy and neglect could follow. Further, no single state is emphasizing education in isolation from others. Rather, all states to one degree or another are focusing on educational "excellence". Thus, major financial and reform efforts in some states may prove to have no consequential effect on the state's relative position in a national ranking.

### 5. State vs. Local Indicators: Practical and Statistical Problems

State level data production runs into a number of practical and political problems. Most notably is the political and philosophical belief that education is a local prerogative and responsibility. Cooperating with a demand for data that will be available on a state to state basis promises little payoff for a local administrator. The data, when aggregated to the state level cannot be attributed to a particular district or school. There is no pride or shame for score production. Lack of investment in the process of data collection has frequently resulted in data that are inaccurate.

Aggregation of data to the state level presents problems of interpretation. Most analysts would suggest that when individual scores are aggregated and considered with group measures such as teacher's education, per pupil cost and so forth, they should be presented separately. This is to avoid a complex correlational analysis implying cause and effect at the individual level. Comber, et. al (1973) present data at the school and individual level on factors such as socioeconomic status, sex and type of school associated with science test results. The percent of variance explained is approximately half as great at the student level as it is for the school level. Socioeconomic status served as a much more effective explanatory variable for school scores (percent explained was 67 percent) than for individual scores (22 percent explained). Others have reported variance for schools in the range of 45 percent and that of individuals 10 to 15 percent (Carter, 1984; White, 1982).

Some argue that interpretations of aggregated data may overestimate the correlation and variation associated with student characteristics. (Robinson, 1950; Slatin, 1974; Bryant el al, 1974). Burstein (1981) argues that aggregation "generally inflates the estimated effects of pupil background and decreases the likelihood of identifying teacher and classroom characteristics that are effective" (p. 195).

Aggregating at the state level ignores the variation that occurs at the district and school level. While many believe there is control within the state on certain indicators, it may be only partial control. For example, within some of the Northwest states, there is a common scope and sequence of curriculum offered as state guides. Districts, however, are allowed to select from many textbook publishers the text and curriculum emphasis they wish to

pursue. This allows fourth graders in one school to be working on pre-basic skills, another school may have fourth graders working on basic skills, and yet another may have them working on higher-order skills.

Using state aggregates requires caution regarding interpretation; school or district units of analysis may require further explanation as well. For example, in matrix sampling, school classrooms are selected to be representative. Yet within some states there is so much in-(California; Texas) or out-(Alaska) migration (Marcus and Mulkeen, 1984) that the school score variation could hardly be attributable to the education in that school, but to the changing composition of the student population.

Using financial measures at the school or district level also clouds the issue. It is a well-established fact that more resources are needed (and in most cases spent) for disadvantaged students (special education teachers, curriculum advisors, aides, etc). Schools in wealthy areas with higher tax bases spend proportionately less on education. Answering the equity issue is not easy when some schools and districts get more money than others and have diverse student populations to serve, and when there is no direct or clear relation between resources and needs.

### 6. The Model and The Measures Potentially Restrict the Collection and Analysis of Information.

The selection of appropriate indicators depends on their purpose or use, and on an appropriate theory or model of education. The purposes of the educational indicators have ranged from describing a state's education program to the development of correlational statistics that imply potential causal

relationships between indicators that affect education. The descriptive approach premise is that the indicator is sufficient in and of itself to reveal important aspects of education. Presenting several indicators in this fashion may allow the reader to deduce empirically or make judgments concerning the indicators and their potential relationships. Others argue for a neat and tidy explanation that might be gained through advanced statistical techniques. These techniques have the aura of scientific truth to the public. Yet, the desire to make their use understandable to the public may mask serious problems. These are discussed below:

The Model of Education. Developing statistical causal relationships requires at a minimum that a model of the educational process be identified. Present technical progress basically supports linear models of elationships. A model based on the familiar economic input-output notion has been generally favored by educators, because technology has not advanced a better one. The economic model can accommodate multiple causation with indicators added incrementally. Models developed by the Chiefs and the Center for Statistics both suggest an input-process-output model. Input includes such variables as expenditures and class size. Process variables may include context and policies. Examples of output variables are test scores and graduation rates. While the models appear to be intuitively descriptive, modern statistical methods can support statistical links as well.

While some would argue that the additive statistical models could never describe the nuances of behavior that go into the educational process, still others would point to the research where traditional measures of input when statistically regressed on output produce equivocal results. Whether intuitively or statistically, then, the links between input and output have

been obscure in the educational realm. Yet, justification for use of this model is made because there are few alternatives. There is a general acceptance of the notion that schools operate like factories, where changes and adjustments are made to the product over time and in the space of the school walls.

Statistical models have become necessary in educational comparisons because of findings that differences, not attributable to the kind of education received, separate the states and districts of this country. Thus, if state A had an average score of 325 on "THE TEST" and state B had an average score of 425, it can not be concluded that these differences are attributed to the content of education alone. This forces those looking for a model to identify appropriate indicators to reduce the differences attributable to factors outside of the education process.

The value of the input-process-output model is that it provides many points at which factors can influence the output. Differences at the input or process level can influence output either directly or indirectly. The critical requirement for the success of the model, however, is to identify those factors at any level of inclusion that reflect how learning really takes place.

The Measures. Selection of indicators follows from the model. Each indicator must be intuitively, politically and statistically important to the model. Measures or indicators that are collected and found to be unrelated to any portion of the model cast doubt on the model or the indicators. With current methods, indicators which repeatedly demonstrate insufficient size of relationship (not merely statistical significance) could be disregarded. Evaluation and assessment staff members of the Northwest state departments of

education have indicated that selection of appropriate indicators is key to the long-term impact on policy making at the state and district levels. If education in one state is found to be "less than average," the indicators should be able to guide the improvement strategies.

Yet, at this point, there seems to be some disagreement on the potential for appropriate indicators to be identified and collected. Of greatest concern is that the indicators as now envisioned represent what can account for the differences at the input level rather than what happens in the content of the educational process. For example, neither of the national projects have identified operational measurements of teacher quality, the content of the curriculum or teaching practice. Measures such as years in teaching, degree status, salary or time spent in inservice training, tell us little of what is going on in the classroom. These latter measures are highly related and may or may not be related to more successful outcomes. Older teachers are rewarded financially for higher degrees and automatically earn larger salaries because of the public system of pay scales. Yet, recent studies of new teachers suggest that the best teachers leave the system within the first five years of teaching (Schlectly and Vance, 1981, Vance and Schlechtly, 1982). If this were the case for the last twenty years of educational history, one could hardly conclude that more experienced, higher salaried teachers are "better" than their younger colleagues.

Another concern is that the measures collected should be ones that reflect the varying organizational choices made in education. If longer school years, experienced teachers or more money per pupil have a large impact on education, then the choices of states and districts are clear. If, however, the greater magnitude of relationship is due to factors outside of educators' control

(socio-economic status, sex, ethnicity, family constellation), then the utility of information on organizational choices is weakened.

Definitions of terms already in use at the state and district level also pose problems for a national indicator project. In the Northwest states, some states collect daily attendance because state funding is dependent on that measure. In other states, funding is based on enrollment with definition and collection of daily attendance varying from district to district.

The definition of dropout or its theoretical converse, graduation rate, is widely variant across the six Northwest states. Most agree that a dropout is one who leaves school prior to completion of the high school course of study and who does not transfer to another school. Length of time from leaving to transfer, counting seriously ill students, students who leave in the spring and do not return in the fall, are all considerations with accompanying variation in the definition. Of significant concern for some districts and at least one Northwest state are the wide variations in enrollment caused by in- or out-migration. If a simple comparison of ninth grade enrollment four years prior is compared to the current 12th grade graduates, these districts or states could have graduation rates of less than fifty percent.

Some states have pointed out the varying definitions in the categories of special needs children. Prompted by lawsuits on the one hand and funding criteria on the other, district and state staff report widely varying counts of learning disabled, mentally handicapped and other special needs children, suggesting that differences in classification present problems.

Fiscal data also suffer for lack of clear definition. Cost per pupil has risen tremendously in the last decade, yet the increase is not believed to have been generally applied at the classroom level. Walberg (1986) points out

that less than one-third of the per pupil cost is attributed to classroom teacher salaries. Thus, when costs have gone up, it has been difficult to relate the increase to classroom results, e.g., achievement gains.

Data on class size are similarly difficult to compile or use for comparisons. How a state handles special education or other categorical program students (self-contained, pull outs, mainstreaming or combinations) affects and obscures class size reports. At best, the organization of special education programs distorts the mean for class size statistics.

Definitional problems, of course, are not unsolvable. Consensus as to definitions and their usage would be needed. Developing such a consensus, however, is seen by some as posing difficulties ranging from the loss of a longitudinal database (based on prior definitions) to requiring changes in state laws.

Measurement technology presents problems as well. Some of the "process" measures identified with the effective schooling literature, and assumed to be important, are the most difficult to operationalize and measure. For example, techniques to measure engaged learning time require training of observers, suffer from generally low reliability rates and introduce "experimenter" effects in the measured classrooms.

The reliability and validity problems associated with self-reported information are well understood by researchers and educators. Student reports suffer from ignorance or embarrassment (e.g., SES of family) to exaggeration (e.g., number of students recalling they took a geometry class). Yet many of the proposed indicators require some self-reporting by students.

Whenever measures are used for rankings, performance or rewards, they are vulnerable to corruption (Campbell, 1979). Instances of school discipline or

vandalism, for example, could easily fall prey to underreporting. Accurate reporting can also be faulty because of the frame of reference of the reporter. One piece of graffiti on a heavily scribbled wall may not seem worth reporting in the mind of one administrator. While, in another school, graffiti on an otherwise clean wall may startle its administrator and seem significant enough to report.

**7. Practical and Logistical Problems Threaten Accuracy**

Issues of sampling, administration and scoring, exclusions and even the perceived burden of any testing and data gathering project impact the desirability of the project and the accuracy of the results. These will be described below:

Test Cycle. Data burden is often related to the time frames for testing. Since many states and districts carry on their own testing programs, additional testing to develop state comparisons will have to fit at times other than when local or state testing is done. Currently, two Northwest states have winter testing and three have fall testing.

Some states have indicated their preference for a national indicator testing cycle of every other year (White Paper, 1986). However, if one assumes that at least three data points are needed to establish a reliable trend, then this option implies that it will be a minimum of four years from the first assessment before a trend can be observed.

Whether the test cycle is fall to spring or fall to fall impacts the results of the testing, as well. Chapter 1 evaluators are familiar with .5 to .7 standard deviation differences in fall to spring testing. But when testing occurs only at each spring--seventh grade spring to eighth grade spring, for example, differences of .1 to .2 standard deviation units occur.

Even order of presentation can affect results. In one state the math test always followed the reading test, with poorer average scores reported for the math testing. One year, the order of the tests was reversed with an increase in math scores and decrease in reading scores (Blust, 1986).

Reporting. It is clear that any national project will have to include procedures and funding to address reporting problems. Delays in reporting of educational statistics have long been a sore point and major source of dissatisfaction for most users. National reporting of results that are three to five years old is easy to disregard. The newsworthiness of the SAT comparisons certainly depends, in part, on their recency.

Sampling Frame. Sampling students in each state/district/school is the current design of the planners for each of the national indicator projects. Sampling has the appearance of a more cost effective and less burdensome procedure for students and test administrators. Whether sampling represents such cost savings over census testing is open to debate for the Northwest states. A survey of evaluation and assessment directors in these states indicated that lack of participation in a state NAEP process was due entirely to costs--in some cases the projected cost greatly exceeded the cost of the census testing of the state.

The issue of sampling may be related to who pays. Currently, large states can receive data from the NAEP surveys indicating "state" results because their population is so large that a sampling is sufficient to provide state results (California and New York). In these cases, the federal government is already paying for state results. Other states may be requested to participate to provide the "state" sampling frame at additional costs.

In a related instance, sampling in states such as Alaska have been attempted with the result that too large a proportion of the population was needed to obtain a sample; thus, as a practical matter, census testing was implemented. In states where there are only 6,000 to 7,000 students per grade level, a sampling frame of 2,000 may reduce the "randomness" of that strategy.

Sampling may be problematic for entirely different reasons as well. Whenever there are mixed levels of aggregation (i.e., individual test scores to school/district or state) and group averages such as median income, percent low income in the area) are collected, there is great potential to distort the resulting analyses. In general, using data generated on the supposed group (e.g., percent poverty in school district) when considered with individually derived data has the result of magnifying that variable in the analysis (Cronbach, 1976; Anderson, 1972). Cronbach (1976) cites the Abt follow-through evaluation where three very different results were shown when data at the pupil, class and school level were presented. In this context, Anderson (1972) suggests that school contextual information (SES, teacher salary) probably will not provide useful information for either theorists or practitioners. He argues that more useful information on context will come, for example, when the variables that distinguish teachers who are paid more are contrasted with those who are paid less.

Part of the problem associated with selecting a sampling frame rests on the theory behind that collection. If a state is the frame of reference, then it is assumed that education is somehow bounded by state factors --- laws, state curriculum, teacher salary schedule, teacher standards and so on. If a district is the sampling frame, then the theory would hold that districts provide education in a more or less bounded way through teacher selection,

curriculum selection, salary and so forth. Similarly, sampling at the school level would suggest the school as the mediating force for achievement. The sampling frame thus represents some theory of the appropriate unit of educational treatment.

A final problem associated with sampling for state or district comparative purposes is the public attitude toward sampling. Most states have gone to census rather than sampling for state assessments, not merely because of the diagnostic nature of the testing for individual students. One evaluation and assessment director pointed out that there is still a public perception that sampling misses too many people. Evidently the Literary Digest's 1936 Alf Landon "victory" is pervasive.

Exclusions. Whether sampling or census, state educational testing was generally allowed for excluding certain classes of students from the testing situation. Exclusions are usually made on the basis that the student is sufficiently handicapped to be unable to adjust to the testing situation. In addition, most states allow discretion at the test site to excuse students for reasons ranging from a recent death in the family to recent immigrant status. This is where the similarity ends. Differences exist in the definition of handicapped--from a student in a self-contained classroom all of the time to one where the student is there a minority of time. New immigrants are excused if they have been in this country for a time period ranging from six months in one state to 18 months or more in another.

Because of the varying degrees of handicapping conditions, it is very difficult to develop with precision procedures that would uniformly be carried out. It could be doubly difficult to gain consensus if the testing purpose was viewed at the local level as one of accountability.

Other classes of students such as private or home-schooled students are excluded/included by state law. If private school students, on average, are even a small percentage above the public school students, then a state or district could be advantaged or not in the resultant rank.

Standardization of Test Procedures. A major implication of plans to date are that the test administration and actual data collection will be decentralized. This differs from a process in which independent data collectors, e.g., outside test administrators, are used. It has obvious advantages in distributing costs to less visible factors, e.g., using existing state and local staff and thereby using less "project" supported funds. However, the extent to which the results will begin to have meaning and engender action within states will depend on the credibility of the administration and collection process. Using local staff has traditionally provided difficulties in ensuring standardization. Thus, there will be pressure to move to external sources if validity and reliability are concerns.

Data Burden. Whenever additional testing or data collection efforts are considered, the issue of data burden is raised. Data burden becomes a policy and technical issue when the data requests are perceived as invasive and not related to the goals or progress of that particular school or district. If a national indicator project uses sampling procedures where district or school level data is not reported back to that district or school, then students, teachers and administrators may not feel responsible for the ultimate outcome. Administrators generally acknowledge that too much testing occurs that is not useful for their school agenda. Anecdotal data from interviewing students who knew they were representatives of a "sample," indicated an almost 10 percent reduction in average scores for 17 year olds (Hill and Kahl, 1986).

Most teachers and administrators feel overwhelmed by the paperwork requirements of the categorical education and civil rights programs of the federal government. As a former Chief State School Officer, Cronin (1986) notes the burden of these requirements on the thousands of smaller districts. He cites the confessions of some district/school officials who reported that press of duties forced them to "guesstimate" numbers when deadlines approached.

Data burden has often been cited by school districts who are routinely selected because of their size or demographic significance in a state. One major district in the South refused to participate in the SREB state comparison project. Negotiations for the inclusion of one northern state were discontinued when only 61 percent of the selected districts agreed to participate (Hill and Kahl, 1986). It is interesting to note that it is sometimes the districts or states with the most complete "local" testing who decline invitations for voluntary participation.

Student burden is also cited as a consideration for participation. Students are tested for their course specific knowledge and may also be tested for categorical, school, district and state testing purposes. Student motivation for participation and performance depends on the perceived value to the student or his school/district/state.

## EXECUTIVE SUMMARY

It is clear that the general public has embraced the notion of a Wall Chart or educational scoreboard comparing states educational programs. Suggestions that such scores are not comparable or must be considered with many qualifications are not viewed as sufficient grounds to deny the public, almost for the first time, access to this information. The public demand for this information is fueled by the general belief that schools have been soft on students and performance has plummeted. The increasing demands for more money are not considered viable any more without a clear indication of what the public is getting for its money.

Educators on the other hand are most conscious of the constraints associated with educational progress. They are acutely aware that the outcome of education is a function of the skills and motivation that the students bring to the classroom, the teacher's ability to transfer and motivate student learning, the curriculum design and the amount of effort expended by the students. Much of the focus of the last several years to improve education for all has rested on the assumption that students were provided unequal access to skilled teachers, well-conceived curriculum and resources. The redirection of resources and structures has not completely reduced the gap for certain students. Thus, it is seen as unfair to compare areas with greater numbers of unprepared students with others.

These cautions have led many educators to seek a fairer method of comparing states or local educational agencies. One method has been to consider a number of indicators of educational achievement; others have been to consider factors that influence outcome and compare only states with

similar types of students. These approaches are being considered by the U.S. Department of Education through its Center for Statistics and within the Council of Chief State School Officers. Each is examining potential indicators for use in comparing school systems and identifying factors associated with success. The implementaion of a national indicator process, however, raises both policy and technical questions which impact the readiness of the states to launch such a project. These issues are summarized:

1. **State or Local Goals for Education May Be Redirected.**

- Policy makers and the general public are more likely to be satisfied with single measures of outcomes. Educators need more extensive information to understand how they can we improve. The information educators need may come from the indicator project, locally developed information or from the results of separate research studies.

- Each state and local district has educational goals based on the needs of its students. An emphasis on basic skills on the one hand, or higher order skills on the other, may redirect the curriculum and teaching toward goals that are too easy or too difficult.

- Setting performance goals may be difficult from both practical and theoretical bases. Those at higher ends of the scale have little room to move while those at the lower end may have an easier time making gains. An overemphasis on meeting goals may encourage corruption of the process and the products.

- The choice of a multiple form test producing a comprehensive profile of learning within a curriculum area will potentially duplicate existing state assessment program goals.

2. **The Potential Costs May Compete with Other Educational Goals**

- The potential costs for any of the national indicator projects are minimally set at $100,000 per state. The upper range has been estimated at $500,000. Given the potentially limited scope of comparative scores and the legitimate demands for state funds, participation in the national indicator project may take money from existing state testing or educational programs. Federal support and /or reduction in costs due to integrating state and national efforts may alleviate the financial burden to the states.

3. **Differences Among Students Confound Educational Outcome Differences**

- The factors associated with differences in educational content are frequently confounded with the differences among the student populations. Adjusting for or separating these into potentially similar groupings suggest that differences among entering students are far more significant than any efforts to remove these differences by the educational process.

4. **Long-Term Utility of the Indicator Process Presents Challenges**

- Developing a global or multi-dimensional measure of educational achievement for each state or district is potentially helpful if it supports reasoned inquiry into the nature of the educational process within the states. However, the potential "box score" could easily change from year to year when compared to others, leaving the analysts uncertain of what to do to improve or maintain, so as not to fall in the rankings.

5. **The Determination of the Unit of Analysis Presents Utility and Practical Problems**

- Considering state averages on achievement measures has the effect of leaving the local educational agency out of the process. While common state laws, teacher standards and funding for standard curriculum would appear to bound each state, the variations among districts or even schools could be even greater. Further, participating in information gathering that is not attributable to one's particular school or district can easily make the process an add-on to an already busy educational agenda.

6. **The Model and the Measures Potentially Restrict the Collection and Analysis of Information**

- The identification of a model of educational process and indicators is intended to simplify and guide understanding. If the measures are collected that are not related to what happens in education, then the analysis obscures rather than clarifies. There are questions in the identification of the appropriate model to portray the educational process as well as with the definitions and assessment of the indicators associated with those models.

7. **Data Burden Poses Problems to Those at the Local Level**

- Many of the states, districts and schools already have numerous testing programs to measure progress and diagnostically identify strengths and weaknesses. Further testing, whether volunteer or mandatory, presents additional burdens to those administering the tests and those taking them. Efforts with few perceived payoffs are likely to be manipulated.

- Using an input-output model of education both constrains and is constrained by related data collection. Observed differences are likely to be population differences rather than educational differences. When population differences are controlled for the differences due to educational alternatives may be small or difficult to disentangle.

- With much local and state testing already going on at the local level, determining the optimal time for an additional testing and data collection effort presents problems. Reporting results in a timely fashion has not easily been accomplished in the past and poses questions of usefulness to those asked to participate.

- Sampling is intended to be a cost effective approach to census testing. Choosing the appropriate level for sampling presents a theoretical issue as it presupposes the boundary of treatment effects. Sampling for state indicators suggests that treatment occurs in the context of state factors, as does sampling at the district represent district factors. Issues associated with varying levels of aggregation pose issues of interpretation.

- Standardization of terms, procedures for testing and collection of data may require changes in state laws and additional expenditures for training and monitoring.

As long as educational accountablility remains at the forefront of the American educational agenda, and as long as obtaining cross-state comparisons of student performance appears to answer the needs of critics and reformers alike, the issues envisioned in implementation can be accommodated by most states. Careful attention to standardization of definitions and collections, installing complementary or integrated testing programs within the states, and providing interpretable feedback on policy implications of the results will be critical to the success of this new program.

## REFERENCES


Anderson, Barry. A Methodological Note on Contextual Effects Studies in Education. Paper presented to Canadian Educational Research Association, 1972.

Blust, R. Comments in Discussion Session, Sixteenth Annual Assessment Conference, June 9-16, 1986, Boulder, Colorado.

Bryant, E.C. et.al. Associations Between Educational Outcomes and Background Variables: A Review of the Literature. National Assessment of Educational Progress Monograph. Denver, 1974.

Burstein, L. and Miller, D. M., Regression-Based Analysis of Multilevel Educational Data in Boruch, R.F., et.al., Reanalyzing Program Evaluations. Josey-Bass, San Franciscok 1981.

Campbell, D., Assessing the Impact of Planned Social Change, Evaluation and Program Planning, 2, 1979.

Cardenas, J. and First, J.M. Children at Risk, Educational Leadership, 1985.

Carter, L.E. The Sustaining Effects Study of Compensatory and Elementary Education. Educational Research, 13, 4-13, 1984.

Coe, Merilyn. Issues and Options for Northwest Regional Educational Laboratory: Databases and Profiling for States and Schools. Northwest Regional Educational Laboratory. March, 1986.

Comber, L.C. and Keeves, J.P., cited in Bryant, E.C. et.al., Associations Between Educational Outcomes and Background Variables: A Review of Selected Literature. National Assessment of Educational Progress Monograph, 1974.

Council of Chief State School Officers. Center on Assessment and Evaluation. Draft Report of Committee on Coordinating Educational Information and Research. October 17, 1985.

Cronbach, L.J. Research on Classrooms and Schools: Formulation of Questions, Design, and Analysis. Occassional Paper of the Stanford Evaluation Consortium, July, 1976.

Hill, R. and Kahl, S. Comments in Discussion Session, Sixteenth Annual Assessment Conference, June 9-16, 1986, Boulder, Colorado.

Marcus, Seldon, and Mulkeen, Thomas, eds. The New Urban Demography: Implication for the Schools. Education and Urban Society, Vol. 16, 4, 1984.

Maudus, G. Testing and Funding: Measurement and Policy Issues, New Directions for Testing and Measurement, 1, 1979.

Moynihan, D. Family and Nation. Harcourt, Brace, Javanovitch, 1986.

Office of Education, Center for Statistics. Plan For The Redesign of the Elementary and Secondary Data Collection Program. March 27, 1986.

Powell, B. and Steelman, L.C. Variations in SAT Performance: Meaningful or Misleading? Harvard Educational Review, 54, 1984.

Robinson, W.S. Ecological Correlates and the Behavior of Individuals. American Sociological Review, 15, 1950.

Schlechtly, P.C. and Vance, V. Do Academically Able Teachers Leave Education? The North Carolina Case. Phi Delta Kappan. 63, 2, 1981.

Seldon, Ramsay. White Paper. Strategies and Issues in the Development of Comparable Indicators for Measuring Student Achievement. State Education Assessment Center, Council of Chief State School Officers. April 30, 1986.

Silverman, L. and Taeuber, R. (eds). Invited Papers: Elementary and Secondary Data Redesign Project. National Center for Educational Statistics. October, 1985.

Slatin, G.T. Ecological Analysis of Delinquency: Aggregation Effects. American Sociological Review, 34, 1969.

The Futures Group. Social Services in the Year 2000. DHHS. 1984.

Vance, V. and Schlectly, P.C. The Distribution of Academic Ability in the Teaching Force: Policy Implications. Phil Delta Kappan, 64, 1982.

Wainer, H. Five Pitfalls Encountered While Trying to Compare States and Their SAT Scores. Journal of Educational Measurement, 23, 1986.

White, Karl. The Relation Between Socioeconomic Status and Academic Achievement. Psychological Bulletin, 91, 1982.

## APPENDIX

### DECISIONS POINTS FOR NATIONAL INDICATOR PROCESS

The decisions identified below will be made either prior to or after the decision by each state to participate in the collection of national indices of educational performance. They are not intended to be exhaustive but to reflect the interrelationships among the decisions.

| DECISIONS | IMPLICATIONS |
|---|---|
| Content of test | Increase in depth and breadth implies either longer testing or multiple forms; increase in cost; greater understanding of curriculum strengths and weaknesses; many scores rather than single score; potential for duplicating existing state assessment programs; changing curriculum scope and sequence; measuring tangential outcomes. |
| Educational model<br>Measures | Choice of model and measures of related variable will restrict the analysis to those choices. Measures are needed that explain differences and are amenable to manipulation. Potential irrelevant factors; may require changes in state laws in defining common measures. |
| Level of aggregation | Will impact interpretation of differences; Support by school and districts |
| Test cycle | Will either interfere or complement existing state and local testing cycles; years in test cycle will effect date when results reveals trends; state laws may have to be changed. |
| Data burden | Related to test cycle, sampling design, test content and state and local testing decisions. |
| Sampling design | Will impact model of education, measures collected; data burden; interpretation of results; costs. |
| Standardization of procedures | Will impact cost; data burden; existing state laws; (e.g., exclusions); training of personnel. |
| Cost | Will depend on test content, scale, sampling frame, test cycle, breadth of information related to outcome; federal assistance. |
| Set performance goals | Some "have no place to go"; other find challenge impossible; Process may be corrupted if viewed as too important. |
| Long term utility | Will determine type of test, measures collected, interpretation, cost |

TABLE 1

NORTHWEST STATE ASSESSMENT CHARACTERISTICS

| State | Test | Subjects | Grade | Cycle | Population | Purposes | Cost (Per student) |
|---|---|---|---|---|---|---|---|
| Alaska | State Developed | Math<br>Reading | 4, 8<br>4, 8 | Bi-annual<br>Winter | Universe<br>@ 15,000 | Curriculum improvement<br>Public accountability<br>Student diagnostic * | $50-60,000<br>($4/student) |
| Hawaii | SAT | Math<br>Reading<br>Writing | 3, 6, 8, 10<br>3, 6, 8, 10<br>3, 6 | Annual<br>6, 8, 10 Fall<br>3 Spring | Universe<br>@ 50,000 | Curriculum improvement<br>Public accountability<br>Student diagnostic * | $200,000<br>($4/student) |
| | CBM | Basic Skills | 3 | | | School level improvement | |
| Idaho | ITBS.,<br>TAP, | Math<br>Reading<br>Writing<br>Science | 8, 11<br>8, 11<br>8, 11<br>8, 11 | Annual<br>Winter | Universe<br>@ 30,000 | Curriculum improvement<br>Public accountability<br>Student diagnostic * | $46,000<br>($1.53/student) |
| | State Developed | Social Stud.<br>Wrtg/ref skls | 8, 11<br>8, 11 | | | | |
| Montana | No state Testing Program | | | | | | |
| Oregon | State Developed | Reading<br>Math<br>Writing | 8<br>8<br>8 | Annual<br>Winter | Sample<br>@ 25,000 | Curriculum improvement<br>Public accountability | $100,00<br>($4/student) |
| Washington | MAT | Math<br>Reading/lang | 4, 8, 10<br>4, 8, 10 | Annual<br>Fall | Universe<br>@ 110,000 | Curriculum improvement<br>Public accountability<br>Student Diagnostic * | $150,000<br>($1.36/student) |

Table drawn from interviews with state Assessment and Evaluation staff, February and June 1986.

*Student diagnostic should be interpreted to include placement or selection for special programs as well as diagnostic information to plan individual students' specific instruction in a classroom.